DOCUMENT RESUME

ED 229 443 TM 830 349

AUTHOR Lance, Charles E.; Moomaw, Michael E.
TITLE Assessing the Psychometric Quality of Performance Rating Scales: Comparisons among Evaluative Criteria.
PUB DATE Mar 83
NOTE 23p.; Paper presented at the Annual Meeting of the Southeastern Psychological Association (Atlanta, GA, March 23-26, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Behavior Rating Scales; *Comparative Analysis; Error of Measurement; *Evaluation Criteria; Item Analysis; Item Banks; *Job Performance; *Psychometrics; Test Construction; Test Format; Testing Problems
IDENTIFIERS *Behaviorally Anchored Rating Scales; Direct Assessment

ABSTRACT

Direct assessments of the accuracy with which raters can use a rating instrument are presented. This study demonstrated how surplus behavioral incidents scaled during the development of Behaviorally Anchored Rating Scales (BARS) can be used effectively in the evaluation of the newly developed scales. Construction of scenarios of hypothetical incumbent job performance and alternative rating instruments makes fuller use of behavioral incident item pools that result from BARS development procedures. Ratee (hypothetical incumbent) performance levels are known from the scale values of items chosen to depict ratee performance and the relative accuracy with which raters may use newly developed BARS can be evaluated in comparison with alternative formats developed as part of the evaluation process. Secondly, the study adds to the literature concerned with comparisons of rating formats in terms of their psychometric properties by contrasting the sole effects of rating format upon the psychometric quality of resulting scales. Again, BARS was an effective format for the rating of the individuals' performance. Finally, the virtue of rating accuracy as an evaluative criterion for assessing the psychometric quality of performance rating scales was extolled. (Author/CM)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED229443

# Assessing the Psychometric Quality of Performance Rating Scales: Comparisons Among Evaluative Criteria

Charles E. Lance

and

Michael E. Moomaw

School of Psychology

Georgia Institute of Technology

Paper Presented to the

Southeastern Psychological Association

March, 1983

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

C. E. Lance

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 830 319

The psychometric properties of performance rating scales have been assessed by the existence of constant rating errors in ratings provided on these instruments. Typically, these assessments are relative ones in that two or more rating formats are compared with respect to their relative psychometric properties (Saal, Downey and Lahey, 1980). Frequently, rating scales have been evaluated along the following rating error criteria: (1) Halo - the tendency to base rating judgments on a global impression of a ratee, or "failure to discriminate among conceptually distinct and potentially independent aspects of a ratee's behavior" (Saal et al., 1980, p. 415); (2) Leniency/Severity - the tendency to assign higher or lower ratings than are warranted by a ratee's performance; and (3) Restriction of Range - truncation of the distribution of ratings compared to that warranted by actual variability in ratees' levels of performance. A rating scale that engenders less of each of these errors, compared to an alternative rating format, is judged to be psychometrically superior.

Conduct of comparative evaluations among rating scales, however, often requires that some rather tenuous assumptions be made regarding several properties of the distribution of "true" levels of employees' performance. For example, one operational definition of halo error examines the magnitudes of intercorrelations among ratings assigned to ratees across performance dimensions. Higher intercorrelations are taken to reflect greater existence of halo error in the ratings. Note the implicit assumption that employee performance levels should not be correlated, or correlations should be low across dimensions (conceptually distinct aspects) of his/her job. That is, there is no consideration of the possibility of the existence of potentially large amounts of "true halo" (Cooper, 1981), or actual covariation among

employees' levels of performance, a plausible condition given a general ability factor (Cronbach and Snow, 1977), for instance. High intercorrelations could also reflect the results of training efforts designed to improve employee skills on those aspects of his/her job where performance was at one time relatively deficient. Thus, it would be possible to erroneously declare one rating scale a "psychometric winner" based on low interdimension correlations while more correct assessments of a high degree of true halo in ratee job performance on an alternative format would lead it to be declared psychometrically inferior.

Similar problems exist for operational definitions of leniency/severity and restriction of range. A common method of assessing leniency involves the calculation of the third moment about the mean, a measure of skewness. The null hypothesis, i.e., the condition of no leniency bias, is that this measure is not significantly different from zero. That is, the assumed underlying true distribution is approximately normal and has a mean located at or very near the scale midpoint. In fact, calculated values of skewness are often significantly negative (cf. Landy, Farr, Saal, & Freytag, 1976). This finding could reflect a leniency error or bias. On the other hand, this finding could reflect the success of selection and promotion programs designed to choose and retain well performing employees. Negatively skewed data could also reflect the effects of employee self-selection, or termination or withdrawal of less successful employees. In short, evaluation of leniency error by assessing degree of skewness in rating data is done against an unknown referent.

The same problem is encountered for a restriction of range criterion. Measures of range restriction (fourth moment about the mean or standard deviation of ratings across ratees within performance dimensions) could also reflect

rating error or the results of organizational influences, e.g., performance norms or ceiling or floor effects. With a rating error criterion correct assignment of cause is impossible.

A final, somewhat related problem exists with another popular evaluative criterion - reliability of ratings. Although several appropriate means for assessing the reliability of a rating instrument exist, the Intraclass Correlation (ICC) (Shrout and Fleiss, 1979) is probably the most popular (Saal et al., 1980). Although an appropriate statistic for the assessment of interrater agreement, any method of calculating the ICC (Shrout and Fleiss, 1979) can seriously underestimate true interrater agreement if there exists low variation among ratees' performance levels (James, Wolf & Demaree, Note 1). Thus, if ratees' performance levels are unknown, so is the accuracy of an ICC estimate of interrater agreement.

In summary, the psychometric quality of rating instruments, implicitly, the accuracy with which raters can use rating instruments, has been inferred from the nonexistence of deviations from some characteristic of an assumed true distribution of employee performance. Since, as a general rule, the population parameters of such a distribution are not known, evaluations or comparative evaluations of scales have been largely conducted with unknown criteria. In this body of literature, Behaviorally Anchored Rating Scales (BARS) (Smith and Kendall, 1963) have received intensive study (Jacobs, Kafry & Zedeck, 1980; Kingstrom and Bass, 1981; Landy & Farr, 1980; Schwab, Heneman & DeCotiis, 1975). Generally, with rating error criteria, BARS have not yielded psychometrically better quality ratings compared to other, often simpler and less expensively developed rating formats (Kingstrom and Bass, 1981; Schwab et al., 1975). Note, however, that this conclusion is reached from research literature that has compared scales in terms of the relative degree to which

rating formats engender rating errors, and thus must be tempered with realization of the ambiguities involved in this mode of comparative evaluation as outlined above.

An alternative set of procedures exists for the evaluation of the psychometric integrity of newly developed rating scales. These are particularly appropriate to the development and evaluation of BARS. The intent of these evaluative methods is to provide (potential) raters with targets (ratees) whose performance effectiveness parameters are identifiable. As will be shown, this is accomplised via the use of scaled behavioral incidents obtained in the process of development of BARS.

The development of BARS involves six general steps: (1) rational definition of performance dimensions; (2) generation of critical incidents of job performance; (3) editing of critical incidents into the form of behavioral expectations; (4) item scaling and "retranslation" of behavioral expectations (Smith and Kendall, 1963); (5) item selection, and (6) final formatting of BARS (see Schwab et al., 1975 for an excellent summary of these procedures). The resulting products of the first four of these steps are a set of rationally defined and consensually unambiguous dimensions of incumbent job performance, along with behavioral expectations scaled as to the degree performance effectiveness represented on a particular job dimension. A behavioral expectation item is eliminated from consideration as a behavioral anchor for BARS if it is not agreed among a criterion percent of judges as to which performance dimension it represents (retranslation criterion) and/or if it is not agreed upon what level of performance effectiveness is represented by the item (standard deviation criterion, DeCotiis, 1978). Once the initial set of behavioral incidents are purged of items thus judged ambiguous, one if left with a pool, now reduced in number, of items deemed suitably unambiguous to qualify as a scale behavioral anchor.

Generally, a larger pool of suitable items results than is required to sufficiently anchor BARS' scales (Zedeck, Jacobs and Kafry, 1976). Writers have recommended using these surplus items to construct parallel forms of BARS (Zedeck, Jacobs and Kafry, 1976) or alternative rating formats (Zedeck, Kafry and Jacobs, 1976). Still others have used these additional items to construct vignettes or scenarios of hypothetical incumbent job performance (DeCotiis, 1977; Sauser, 1979). It is in this final application of these surplus items that the potential for alternative scale evaluation procedures lie.

Incorporation of scaled critical incidents in a narrative description of employee performance provides a means by which levels of employees' performance effectiveness can be specified, i.e., by the scale values chosen to depict performance. While relative degree of rating accuracy has been inferred from the relative absence of traditional errors in rating data, when true performance levels are known, rating accuracy can be assessed directly by a simple metric: deviation of rating from true score, or performance level as depicted. Thus, direct assessments of the accuracy with which raters can use a rating instrument are made available.

One purpose of the present study was illustrative: to demonstrate possible extended uses of surplus, psychometrically acceptable behavioral expectation items. Another secondary purpose was to evaluate an alternative rating format according to its psychometric efficacy in comparison with a BARS. Primarily, this study evaluated the usefulness of an alternative criterion for the evaluation of the psychometric properties of rating scales: direct assessment of rating accuracy.

## METHOD

### Construction of Experimental Materials

As a result of prior interview and task inventory approaches to job

analysis of a set of selected secretarial jobs, eleven dimensions of secretarial job performance were rationally defined (York, Note 2). Next, positive and negative critical incidents of job performance were solicited from a sample of secretarial incumbents. Obtained incidents were edited to yield 500 brief evaluative statements relevant to secretarial job performance. During editing, care was taken to preserve as much of incumbents' own language and terminology as possible. These 500 statements were randomly assigned to and randomly arranged within five sets of 100 statements each. Another sample of secretaries (n = 100) then participated in item scaling by Thurstone's Method of Equal-Appearing Intervals (Edwards, 1957) and retranslation of statements to performance dimensions as defined. Each subject was randomly assigned one booklet containing 100 statements. Subjects rated each statement on a 7-point scale as to the level of performance exemplified, and allocated each statement to one of the rationally defined performance dimensions. The following criteria were adopted to assure that only unambiguous behavioral items would be retained as potential BARS anchors. An item was retained: (1) if it had a Q-value of less than 1.90 (Edwards, 1957); and (2) if it was allocated to one performance dimension by at least 67% of the respondees, with the constraint that no more than 20% of the other responses fell into any other one category. A total of 208 items met these criteria. An insufficient number of items were retranslated to two performance dimensions to adequately anchor scales for them, thus these were eliminated from further consideration in this study. The remaining nine performance dimensions are listed in Table 1.

Behaviorally Anchored Rating Scales Between four and six behavioral items with larger percentage retranslations and smaller Q-values were selected to anchor scales for each performance dimension at points as evenly spaced along the scale as possible. The final BARS included the name of each performance dimension, a definition of that dimension, a vertical 7-point graphic scale

anchored numerically and adjectivally and the behavioral anchors located along the sides of the graphic scale with an arrow pointing to a point near that corresponding to the anchor's scale value.

Weighted Checklist Items for an alternative Weighted Checklist (WCL) rating format were selected from the common item pool and according to the same judgmental criteria as for the BARS. Five items for each of the nine performance dimensions were selected whose scale values represented as broadly and as evenly as possible the entire range of the 7-point scale. The final format included only the 45 items arranged, randomly, following instructions for scale use. Use of the scale involved raters endorsing those items judged to be descriptive of a ratee's typical job performance.

Scenarios of Secretarial Job Behavior Again, from the same common item pool, two items each were selected to describe hypothetical incumbent performance on two different scenarios. These two items reflected nearly the same level of performance within each job performance dimension (i.e., nearly equal scale values), either a high or low performance level on one scenario. Two items from the opposite end of the scale continuum were selected for the other scenario. This arrangement is depicted in Table 2. The selected items were randomly arranged and formatted to follow a brief description of a hypothetical secretary arbitrarily named either "Cathy" or "Debra".

Procedure

Seventy-five secretaries were randomly assigned to one of two groups. Subjects in each group rated the performance of one of the hypothetical incumbents on both rating formats. All correspondence was conducted by mail and complete confidentiality of responses to the investigators was assured.

## RESULTS

### Rating Errors

Halo Contradictory results were obtained regarding format superiority under different operational definitions of halo employed in the literature (Saal et al., 1980). First, intercorrelations of ratings across job performance dimensions were significantly larger for the WCL than for the BARS, indicating less halo error on the BARS. Second, Principal Components Analyses (Mulaik, 1972) yielded two components that accounted for approximately 75% of the rating variance for each format/scenario combination, indicating no difference in halo error between the two rating formats. Third, standard deviations of each rater's ratings of one scenario across performance dimensions were significantly larger for the WCL than for the BARS, indicating less halo error in ratings obtained on the WCL format. Thus, the first and third commonly used operational measures of halo error suggested opposite conclusions regarding the relative existence of halo error on the two rating formats.

Leniency/Severity One statistical test often employed to assess the existing leniency or severity error is to test for a difference between the mean of obtained ratings and the scale midpoint, the theoretical population mean (Saal et al, 1980). Of 18 such comparisons (two scenarios by nine performance dimensions each), 13 mean BARS ratings differed significantly from the scale midpoint and ten significant differences were obtained for the WCL. Without exception, the differences in means from the scale midpoints were in the direction of depicted performance level (i.e., high or low performance effectiveness). Without the knowledge of actual performance levels, however, one might conclude that both rating formats engendered wild variations in ratings across performance dimensions. Similar results were obtained for another operational definition of leniency error, the third moment about the mean (skewness).

Though negative in the majority, both significantly positive and negative values of skewness were obtained for both rating formats. These findings parallel the ones above and could lead one to conclude that either raters cannot use scales appropriately or that somehow the scale midpoints of various dimensions' scales have been mislocated.

Restriction of Range Even greater confusion was generated with two different statistical measures of range restriction. First, standard deviations of ratings within performance dimensions (i.e., across raters) indicated that BARS almost invariably exhibited greater restriction of range (i.e., smaller standard deviations) than the WCL, but only on those performance dimensions wherein low performance effectiveness was depicted. Where high performance was depicted there were no differences. An entirely different picture was painted with a second operational definition of range restriction, the fourth moment about the mean (kurtosis). Significantly positive kurtosis values indicate a distribution more widely dispersed than normal (platykurtosis) and negative values reflect narrower dispersion (leptokurtosis). No values of kurtosis were significantly different from zero for the BARS data. For the WCL data, on the other hand, in six of nine instance where high performance was depicted the data were significantly platykurtic and in eight of nine instance where low performance was depicted the data were significantly leptokurtic. Thus, these two sets of results are not even remotely convergent.

Evaluation of Rating Error Criteria Clearly, the two ratees (scenarios) in the present study are not apt to be representative of secretaries' configurations of their job performance effectiveness across various aspects (dimensions) of their job. Secretaries are unlikely to perform excellently on exactly half of their job duties and extremely poorly on the other half. The data presented here, however, illustrate an extreme case of what can happen when rating error

criteria are used to evaluate rating scales when there is variation either among ratees or within ratees across dimensions of job performance. The above results demonstrate the equivocalness of the conclusions that might be drawn from such an evaluation study. Either rating format could have been implicated as engendering greater halo error depending upon the statistical definition chosen, and virtually no clear conclusions could be drawn from assessments of relative range restriction or leniency bias, although ratings on the WCL were somewhat more greatly elevated overall. In general, rating error criteria are not recommended.

Accuracy

As noted above, in the construction of scenarios of secretarial job performance, it was possible to determine a priori the level of job performance effectiveness that would be depicted in each scenario by choosing scaled behavioral incidents, similar in scale value, to represent high or low performance levels across job dimensions. Thus, as depicted, performance effectiveness levels of ratees across performance dimensions were known. Given this, assessment of the accuracy with which raters can use alternative rating scales is straightforward. One need only quantify deviations of ratings from known performance levels. Measures of rating inaccuracy were calculated for both formats on each performance dimension as average squared deviations from true (known) performance levels:

$$\overline{Acc}_j = \frac{1}{n} \sum_{i=1}^{n} (X_{ij} - T_j)^2$$

where $\overline{Acc}_j$ is the mean rating inaccuracy for the $j$th performance dimension, $X_{ij}$ is the $i$th rater's rating of ratee performance on dimension $j$, and $T_j$ is the is the performance level depicted on that dimension. Results of Wilcoxon sign-rank tests for differences between formats are presented in Table 3. Recall

that what is presented are median rating inaccuracy scores: smaller values indicate more accurate ratings. Note that there are few differences between formats on those dimensions wherein high performance was depicted. This is explained by the generally elevated ratings on the WCL. Overwhelmingly, however, raters provided more accurate evaluations of poor performance on the BARS format. This is particularly important in light of the general finding of elevation of ratings (Kingstrom and Bass, 1981). An accepted finding is that ratings completed for administrative purposes are more lenient than ratings completed anonymously for research purposes (Landy and Farr, 1980). Motivational factors aside, the present results suggest that raters can rate poor performance more accurately on a BARS than an alternative WCL rating format. This is, raters were better able to assign an accurately discriminated evaluation of poorer performance on the BARS, while evaluations of more effective performance were approximately equally assigned on both the BARS and WCL. Thus, the comparative criterion of rating inaccuracy, a measure that directly assesses validity of ratings, suggested BARS as the superior rating format.

Reliability

In addition to validity, reliability represents another important criterion for criteria. Note that reliability is a necessary but not sufficient condition for valid ratings, and conversely, validity is sufficient but not necessary for reliability. The present experimental design readily lends itself to assessment of interrater reliability via the Intraclass Correlation (ICC). As mentioned earlier, lack of between-ratee differences in actual performance can lead to erroneous estimates of interrater reliability. The application of the ICC as an index of interrater reliability in the present case is appropriate since, as described above, scenarios were constructed to reflect varying degrees of effectiveness of performance across performance dimensions. As shown in Table

4, ratings were generally more reliable on the BARS than on the WCL (median ICC's = .42 and .22 respectively).

## DISCUSSION

In line with stated objectives, the present study addressed three main issues. First, it was demonstrated how surplus behavioral incidents scaled during the development of BARS can be used effectively in the evaluation of the newly developed scales. Construction of scenarios of hypothetical incumbent job performance and alternative rating instruments makes fuller use of behavioral incident item pools that result from BARS development procedures. Ratee (hypothetical incumbent) performance levels are known from the scale values of items chosen to depict ratee performance and the relative accuracy with which raters may use newly developed BARS can be evaluated in comparison with alternative formats developed as part of the evaluation process.

Secondly, the present study adds to the already large body of literature concerned with comparisons of rating formats in terms of their psychometric properties. In the past, researchers have often confounded rating format, developmental procedures and job performance domain surveyed by the rating scale in their comparisons among instruments (cf., Borman and Dunnette, 1975; Burnaska and Hollman, 1974; DeCotiis, 1977). The present study, as have some others (cf. Zedeck, Kafry and Jacobs, 1976) contrasted the sole effects of rating format upon the psychometric quality of resulting scales. Again, BARS was supported as an effective format for the rating of individuals' performance.

Finally, the virtue of rating accuracy as an evaluative criterion for assessing the psychometric quality of performance rating scales was extolled. The use of the metric of rating inaccuracy described here, however, assumes that some more objective measure of what quality is being rated is available. The idea of using standardized stimuli as ratees is not new. DeCotiis (1977) and

Sauser (1979) both constructed standardized scenarios of incumbent job performance much in the manner of the present study. Also, Borman (1979) used videotaped job performance as the rating stimulus. The recommendation here is for more routine use of standardized stimuli such as vignettes of performance for the evaluation of rating instruments. Also, the procedures recommended here are not limited to construction of BARS, although they are most applicable here. Similar procedures could be adapted for Likert-type scale development.

The primary advantage of evaluating performance rating instruments in terms of the accuracy with which raters can use the scales lies in the directness of the approach. As discussed above, rating error criteria are attempts to quantify deviations from accurate ratings indirectly. On the other hand, a metric of deviations of rated values from true performance scores such as the one utilized in the present study direction assess rating accuracy.

One obvious limitation to the approach advocated here is its generalizability to actual use, for example, the rating of real individuals active in ongoing organizational activities. That is, results from procedures such as those outlined in this paper may not be strongly externally valid. On the other hand, these procedures may represent the strongest instance of internal validity. Rating of vignettes of hypothetical incumbent job performance may be conditions conducive to the most accurate possible use of newly developed rating instruments. These ideal conditions will simply not exist in the "real world".

The complete evaluation of newly developed rating instruments may inevitably require a two-step process. Primarily one may wish to assess the accuracy with which raters can evaluate ratee performance with the use of stnadardized stimuli such as job performance scenarios. Secondarily, pilot testing of scales with supervisory ratings of subordinate performance, using rating error

criteria, would hopefully reinforce the conclusions from the prior analysis. In such a secondary analysis, however, the researcher need be aware of the assumptions necessarily made with rating error criteria, and interpret the results of such an evaluation study in an appropriately circumspect manner. So done, these two approaches to the evaluation of the psychometric properties of performance evaluation instruments can be complementary.

## Reference Notes

1. James, L.R., Wolf, G., & Demaree, R.G. Estimating interrater reliability in incomplete designs. (Office of Naval Research Tech. Rep. IBR 81-14). Institute of Behavioral Research, Texas Christian University, Fort Worth, TX, August, 1981.

2. York, C.M. Clerical selection validation study. Project Interim Report. Atlanta: Georgia Institute of Technology, February, 1981.

## References

Borman, W.C. Format and training effects on rating accuracy and rating errors. Journal of Applied Psychology, 1979, 64, 410-421.

Borman, W.C. & Dunnette, M.D. Behavior-based versus trait-oriented performance ratings: An empirical study. Journal of Applied Psychology, 1975, 60, 561-565.

Burnaska, R.F. & Hollman, T.D. An empirical comparison of the relative effects of rater response biases on three rating scale formats. Journal of Applied Psychology, 1974, 59, 307-312.

Cooper, W.H. Ubiquitous halo. Psychological Bulletin, 1981, 90, 218-244.

Cronbach, L.J. & Snow, R.E. Aptitudes and Instructional Methods. NY: Irvington Publishers, Inc., 1977.

DeCotiis, T.A. An analysis of the external validity and applied relevance of three rating formats. Organizational Behavior and Human Performance, 1977, 19, 247-266.

DeCotiis, T.A. A critique and suggested revision of behaviorally anchored rating scales developmental procedures. Educational and Psychological Measurement, 1978, 38, 681-691.

Edward, A.L. Techniques of Attitude Scale Construction. NY: Appleton-Century-Crofts, 1957.

Jacobs, R., Kafry, D. & Zedeck, S. Expectations of behaviorally anchored rating scales. Personnel Psychology, 1980, 33, 595-640.

Kingstrom, P.O. & Bass. A.R. A critical analysis of studies comparing behaviorally anchored rating scales (BARS) and other rating formats. Personnel Psychology, 1981, 34, 263-289.

Landy, F.J. & Farr, J.L. Performance rating. Psychological Bulletin, 1981, 87, 72-107.

Landy, F.J., Farr, J.L., Saal, F.G., & Freytag, W.R. Behaviorally anchored rating scales for rating the performance of police officers. Journal of Applied Psychology, 1976, 61, 752-758.

Mulaik, S.A. The Foundations of Factor Analysis. NY: McGraw-Hill, 1972.

Saal, F.E., Downey, R.G., & Lahey, M.A. Rating the ratings: Assessing the psychometric quality of rating data. Psychological Bulletin, 1980, 88, 413-428.

Sauser, W.I. A comparative evaluation of the effects of rater participation and rater training on characteristics of employee performance appraisal ratings and related mediating variables. Unpublished doctoral dissertation, Georgia Institute of Technology, Atlanta, GA, 1978.

Schwab, D.P., Heneman, H.G., III, & DeCotiis, T. Behaviorally anchored rating scales: A review of the literature. Personnel Psychology, 1975, 28, 549-562.

Shrout, P.E. & Fleiss, J.L. Intraclass correlations: Uses in assessing inter-rater reliability. Psychological Bulletin, 1979, 86, 420-428.

Smith, P.C. & Kendall, L.M. Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. Journal of Applied Psychology, 1963, 47, 149-155.

Zedeck, S., Jacobs, R., & Kafry, D. Behavioral expectations: Development of parallel forms and analysis of scale assumptions. Journal of Applied Psychology, 1976, 61, 112-115.

Zedeck, S., Kafry, D., & Jacobs, R. Format and scoring variations in behavioral expectation evaluations. Organizational Behavior and Human Performance, 1976, 17, 171-184.

Table 1

Secretarial Job Performance Dimensions

A. Bookkeeping and Financial

B. Composing or Editing

C. Filing, Sorting, Routing, etc.

D. Gathering Information

E. Handling Materials

F. Communications and Public Relations

G. Operating or Maintaining Machines

H. Supervising, Directing, Deciding

I. Typing or Data Entry

Table 2

Scenarios of Secretarial Job Performance*

| Performance Dimension | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| Scenario | | | | | | | | | |
| "Cathy" | L | H | L | H | H | L | H | L | H |
| "Debra" | H | L | H | L | L | H | L | H | L |

*Performance effectiveness depcited is either High (H) or Low (L)

Table 3

Comparisons Among Formats' Median Rating Inaccuracy Scores

| Scenario 1 | | | | Scenario 2 | | | |
|---|---|---|---|---|---|---|---|
| Perf. Dimen.[1] | BARS | WCI | T[2] | Perf. Dimen. | BARS | SCL | T |
| A(L) | 1.50 | 3.82 | 197** | A(H) | 2.28 | 2.95 | 179.5*** |
| B(H) | .49 | 1.10 | 264.5 | B(L) | 1.98 | 4.98 | 108*** |
| C(L) | 2.76 | 1.61 | 268 | C(H) | .21 | .58 | 250 |
| D(H) | 1.80 | 2.53 | 266 | D(L) | 1.28 | 9.49 | 111.5*** |
| E(H) | 1.40 | 2.19 | 338 | E(L) | 4.20 | 21.52 | 47*** |
| F(L) | 2.48 | 1.88 | 282.5 | F(H) | .36 | 1.17 | 279 |
| G(H) | .57 | 1.99 | 197.4** | G(L) | 2.24 | 15.55 | 115*** |
| H(L) | .48 | 3.03 | 101*** | H(H) | 3.20 | .79 | 176.5*** |
| I(H) | .55 | 2.37 | 262 | I(L) | 2.02 | 11.70 | 127*** |

[1]Depicted performance is either high (H) or low (L)

[2]Wilcoxon T statistic for Rank-Sign test

** p less than .02

*** p less than .01

Table 4

Intraclass Correlations

| Performance Dimension | BARS | WCL |
|---|---|---|
| A | .425 | .485 |
| B | .471 | .234 |
| C | .485 | .479 |
| D | .242 | .000 |
| E | .180 | .000 |
| F | .419 | .513 |
| G | .346 | .182 |
| H | .460 | .363 |
| I | .320 | .178 |